Question Answering for Machine Reading Evaluation

Álvaro Rodrigo[1], Anselmo Peñas[1],
Eduard Hovy[2] and Emanuele Pianta[3]

[1]NLP & IR Group, UNED, Madrid {alvarory,anselmo@lsi.uned.es}
[2] USC-ISI {hovy@isi.edu}
[3] CELCT {pianta@fbk.eu}

Abstract. Question Answering (QA) evaluation potentially provides a way to evaluate systems that attempt to understand texts automatically. Although current QA technologies are still unable to answer complex questions that require deep inference, we believe QA evaluation techniques must be adapted to drive QA research in the direction of deeper understanding of texts. In this paper we propose such evolution by suggesting an evaluation methodology focused on the understanding of individual documents at a deeper level.

Keywords: Question Answering, Machine Reading, Evaluation

1 Introduction

Question Answering (QA) evaluations measure the performance of systems that seek to "understand" texts. However, this understanding has so far been evaluated using simple questions that require almost no inferences to find the correct answers. These surface-level evaluations have promoted QA architectures based on Information Retrieval (IR) techniques, in which the final answer(s) is/are obtained after focusing on selected portions of retrieved documents and matching sentence fragments or sentence parse trees. No real understanding of documents is performed, since none is required by the evaluation.

Other evaluation tasks have proposed a deeper analysis of texts. These include the Recognizing Textual Entailment (RTE) Challenges[1], the Answer Validation Exercise (AVE)[2], and the pilot tasks proposed at the last RTE Challenges[3].

Recently, Machine Reading (MR) has been defined as a new version of an old challenge for NLP. This task requires the automatic understanding of texts at a deeper level [4]. The objective of an MR system is to extract the knowledge contained in texts for improving the performance of systems in tasks that involves some kind of reasoning. However, there is not yet a clear evaluation strategy for MR systems.

Given that MR systems use the knowledge of texts in reasoning tasks, in this

[1] http://pascallin.ecs.soton.ac.uk/Challenges/RTE/
[2] http://nlp.uned.es/clef-qa/ave/
[3] http://www.nist.gov/tac/2010/RTE/index.html

paper we propose to evolve QA evaluations in order to evaluate MR systems. That is, we propose to design an evaluation methodology for MR systems which takes advantage of the experience obtained in QA, RTE, and AVE evaluations, with the objective of evaluating a task where a deeper level of inference is required.

The paper is structured as follows: Section 2 proposes an evaluation methodology of MR systems. Section 3 describes previous evaluations related to the one proposed in this paper. Finally, some conclusions are given in Section 4.

2 Evaluation Proposal

Objective. The objective of an MR system is to understand the knowledge contained in texts in order to improve the performance of systems carrying out reasoning tasks. MR is a task strongly connected with Natural Language Processing (NLP). Given this connection between NLP and MR, it is expected that MR can benefit from the experience obtained in more than 50 years of NLP research. We think the current state of NLP technologies offers a good opportunity for proposing an evaluation of MR systems.

We propose an evaluation approach that represents an evolution of previous NLP evaluations, which are described in Section 3.

Form of test. In the evaluation proposed here, systems would receive a set of documents and a questionnaire for each document. Each questionnaire would be used for checking the understanding of a single document. Thus, the typical IR step of finding relevant documents is not required, and the system can focus on understanding the document.

The evaluation would measure the quality of a system's answers to the questions for each document. The objective of a system is to pass the test associated to each document, indicating that the system understood the document. The final evaluation measure would count the number of documents that have been understood by a system.

This evaluation requires systems to understand the test questions, analyze the relation among entities contained in questions and entities expressed by candidate answers, and understand the information contained in documents. An answer must be selected for each question, and the reasoning behind the answer must be given. Therefore, it is a task in which different NLP tasks converge, including QA, RTE, and Answer Validation (AV).

Question content. A series of increasingly sophisticated questions, and increasingly challenging answer forms, can be developed over the years. Question evolution can for example proceed as follows:

- simple factoids: facts that (as in traditional QA evaluation) are explicitly present in the text
- facts that are explicitly present but are not explicitly related (for example, they do not appear in the same sentence, although any human would understand they are connected)

- facts that are not explicitly mentioned in the text, but that are one inferential step away (as in the RTE challenge)
- facts that are explicitly mentioned in the text but that require some inference to be connected to form the answer
- facts that are not explicitly mentioned in the text and that require some inference to be connected to form the answer

Answer forms. Different answer types can be suggested. We propose to begin with multiple-choice tests, where a list of possible answers for each question is given and the system has to select only one of the given answers. As systems increase their capabilities, the answer form can evolve from multiple-choice tests to cloze tests to open-ended answer formulation tests to task performance tests.

An example multiple-choice test is shown in Fig. 1 (the question is about a text of junk food). This evaluation is similar to tests for people learning a new language, whose reading comprehension is checked using different tests that measure their understanding of what they are reading. Hence, we see our evaluation as a test where systems obtain marks that represent their understanding of texts. These tests can be of different difficulty depending on the understanding level required.

According to the article, some parents:

A. tend to overfeed their children
B. believe their children don't need as many vitamins as adults
C. claim their children should choose what to eat
D. regard their children's bad eating habits as a passing phase

Fig. 1. Example multiple choice question

Another popular test for language learners is the cloze test, in which the learner (in our case: the system) has to fill one or more words into a given sentence or phrase. The phrase is carefully constructed so that a more-than-superficial reading is required in order to fill the gap correctly. Continuing the example in Fig. 2.

According to the article, some parents teach their children bad eating habits by _____ .

Fig. 2. Example cloze question

Test procedure. Given the fact that systems might contain built-in background knowledge, it is important to determine as baseline how much a system knows before it reads the given text. We therefore propose to apply the test twice for each text:

- First, the system tries to answer the questions without having seen any text;
- Second, the system reads the text;
- Third, the system answers the same questions.

The system's score on the first trial (before it has seen the text) is subtracted from the system's score on the second.

Domain. Regarding document collections, we suggest using documents from different topics. For this reason we propose to perform the evaluation over world news. This domain covers several categories that contain different phenomena. Therefore, these documents offer the possibility of evaluating general-purpose systems.

3 Related Work

The evaluation proposed in this paper is an extension of previous evaluations of automatic NLP systems. The first of these evaluations is QA, where a system receives questions formulated in natural language and returns one or more exact answers to these questions, possibly with the locations from which the answers were drawn as justification [2]. The evaluation of QA systems began at the eighth edition of the Text Retrieval Conference (TREC)[4] [6], and has continued in other editions of TREC, at the Cross Language Evaluation Forum (CLEF)[5] in the EU, and at the NII-NACSIS Test Collection for IR Systems (NTCIR)[6] in Japan.

Most of the questions used in these evaluations ask about facts (as for example Who is the president of XYZ?) or definitions (for instance What does XYZ mean?). These questions do not require the application of inferences or even a deeper-than-surface-parse analysis for finding correct answers. Besides, since systems could search for answers among several documents (using IR engines), it was generally possible to find in some document a "system-friendly" statement that contained exactly the answer information stated in an easily matched form. This made QA both shallow and relatively easy. In contrast, by giving only a single document per test, our evaluation requires systems to try to understand every statement, no matter how obscure it might be, and to try to form connections across statement in case the answer is spread over more than one sentence.

On the other hand, our evaluation benefits from past research in QA systems and QA-based evaluations, specifically in the analysis and classification of questions, different ways of evaluating different QA behaviors, etc.

The following related evaluation is the Recognizing of Textual Entailment (RTE), where a system must decide whether the meaning of a text (the Text T) entails the meaning of another text (the Hypothesis H): whether the meaning of the hypothesis can be inferred from the meaning of the text [3].

RTE systems have been evaluated at the RTE Challenges, whose first edition was proposed in 2005. The RTE Challenges encourage the development of systems that have to treat different semantic phenomena. Each participant system at the RTE Challenges received a set of text-hypothesis (T-H) pairs and had to decide for each T-H pair whether T entails H.

[4] http://trec.nist.gov/
[5] http://www.clef-campaign.org/
[6] http://research.nii.ac.jp/ntcir/

These evaluations are more focused on the understanding of texts than QA because they evaluate whether the knowledge contained in a text imply the knowledge contained in another. Then, our evaluation would benefit from the RTE background in the management of knowledge.

Our evaluation differs from RTE because RTE is a simple classification task (either T entails H or it does not), whereas we require extracting the knowledge that answers a question, not for checking whether the text is contained in another text.

A combination of QA and RTE evaluations was done in the Answer Validation Exercise (AVE) [8,9,10]. Answer Validation (AV) is the task of deciding, given a question and an answer from a QA system, whether the answer is correct or not. AVE was a task focused on the evaluation of AV systems and it was defined as a problem of RTE in order to promote a deeper analysis in QA.

Our evaluation has some similarities with AVE. The multiple choice test we propose can be approached with an AV system that selects the answer with more chances of being correct. However, we would give as support a whole document while in AVE only a short snippet was used. Besides, AVE used questions defined in the QA task at CLEF, which were simpler (they required less inference and analysis) than the ones we propose.

The proposal of ResPubliQA 2009 at CLEF [5] had the objective of transferring the lessons learned at AVE to QA systems. With this purpose, ResPubliQA allowed to leave a question unanswered in case of a system was not sure about finding a correct answer to that question. The objective was to reduce the amount of incorrect answers while keeping the number of correct ones, by leaving some questions unanswered. Thus, it was promoted the use of AV modules for deciding whether to ask or not a question.

Another application of RTE, similar to AVE, in the context of Information Extraction is going to be made in a pilot task defined at the RTE-6[7] with the aim of studying the impact of RTE systems in Knowledge Base Population (KBP)[8]. The objective of this pilot task is to validate the output of participant systems at the KBP slot filling task that was celebrated at the Text Analysis Conference (TAC)[9].

Systems participating at the KBP slot filling task must extract from documents some values for a set of attributes of a certain entity. Given the output of participant systems at KBP, the RTE KBP validation pilot consists of deciding whether each of the values detected for an entity is correct according to the supporting document. For taking this decision, participant systems at the RTE KBP validation pilot receive a set of T-H pairs, where the hypothesis is built combining an entity, an attribute and a value.

This task is similar to the one proposed here because it checks the correctness of a set of facts extracted from a document. However, the KBP facts are very simple because they ask about properties of an entity, whereas the QA evaluation we propose can in principle ask about anything. Therefore, our task would, as it evolves, require a deeper level of inference.

[7] http://www.nist.gov/tac/2010/RTE/index.html
[8] http://nlp.cs.qc.cuny.edu/kbp/2010/
[9] http://www.nist.gov/tac/2010/

Finally, we want to remark that there have been other efforts closer to our proposal for evaluating understanding systems, as the "ANLP/NAACL 2000 Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems[10]".

This workshop proposed to evaluate understanding systems by means of Reading Comprehension (RC) tests. These tests are similar to the ones suggested in this paper. That is, the evaluation consisted of a set of texts and a series of questions about each text. Although the approach may have not changed, the field has now made many steps forward and we think that the current state of systems is more appropriate for suggesting this evaluation. In fact, most of the approaches presented at that workshop showed how to adapt QA systems to such kind of evaluation.

A more complete evaluation methodology of MR systems has been reported in [7], where the authors proposed to use also RC tests. However, the objective of these tests was to extract correct answers from documents, what is similar to QA without an IR engine. In our evaluation, we would ask for selecting a correct answer from a set of candidate ones, where the correct answer contains a knowledge that is present in the document, but this knowledge is written in a different way. Thus, we ask for a better understanding of documents than in RC tests.

4 Conclusions

Current research on NLP technologies has gradually led to systems that may now attempt a deeper-than-surface understanding of texts. A series of evaluations of previous systems has allowed these advances. However, a new evaluation is required to drive the increasing deepening of understanding and use of inference to augment surface-level and deeper structure matching. We propose here an evaluation using question answering on single documents, where the answers require increasingly deep levels of inference. This evaluation moves effort away from retrieval and toward reasoning, which is a prerequisite for true text understanding.

Acknowledgments.

This work has been partially supported by The Spanish Government through the "Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de I+D+i" 2008-2011 (Grant PR2009-0020), the Spanish Ministry of Science and Innovation within the project QEAVis-Catiex (TIN2007-67581-C02-01), the Regional Government of Madrid under the Research Network MA2VICMR (S-2009/TIC-1542), the Education Council of the Regional Government of Madrid, the European Social Fund and the US Advanced Defense Research Programs Agency DARPA, under contract number FA8750-09-C-0172.

[10] http://www.aclweb.org/anthology/W/W00/#0600

References

1. Eric J. Breck, John D. Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light, and Inderjeet Mani. How to Evaluate your Question Answering System Every Day and Still Get Real Work Done. In Proceedings of the Second Conference on Language Resources and Evaluation (LREC-2000), pages 1495–1500, 2000.
2. Eric Brill, Jimmy J. Lin, Michele Banko, Susan T. Dumais, and Andrew Y. Ng. Data-Intensive Question Answering. In Proceedings of the Tenth Text REtrieval Conference (TREC), 2001.
3. Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Lecture Notes in Computer Science, volume 3944, pages 177–190. Springer, 2005.
4. Oren Etzioni, Michele Banko, and Michael J. Cafarella. Machine reading. In Proceedings of the 21st National Conference on Artificial Intelligence, 2006.
5. Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forascu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, Petya Osenova. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In C. Peters, G. di Nunzio, M. Kurimo, Th. Mandl, D. Mostefa, A. Peñas, G. Roda (Eds.), Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments, Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, 30 September - 2 October. Revised Selected Papers. (to be published)
6. Ellen M. Voorhees and Dawn M. Tice. The TREC-8 Question Answering Track Evaluation. In Text Retrieval Conference TREC-8, pages 83–105, 1999.
7. B. Wellner, L. Ferro, W. Greiff and L. Hirschman.. Reading comprehension tests for computer-based understanding evaluation. Nat. Lang. Eng. 12, 4, 305-334. 2006
8. Anselmo Peñas, Álvaro Rodrigo, Felisa Verdejo. Overview of the Answer Validation Exercise 2007. In C. Peters, V. Jijkoun, Th. Mandl, H. Müller, D.W. Oard, A. Peñas, V. Petras, and D. Santos, (Eds.): Advances in Multilingual and Multimodal Information Retrieval, LNCS 5152, September 2008.
9. Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, Felisa Verdejo. Overview of the Answer Validation Exercise 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (Eds.): Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers.
10. Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo. Overview of the Answer Validation Exercise 2008. In C. Peters, Th. Mandl, V. Petras, A. Peñas, H. Müller, D. Oard, V. Jijkoun, D. Santos (Eds), Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers.